

INFERRING THE CONDITIONAL MEAN

GUSZTÁV MORVAI AND BENJAMIN WEISS

ABSTRACT. Consider a stationary real-valued time series $\{X_n\}_{n=0}^\infty$ with a priori unknown distribution. The goal is to estimate the conditional expectation $E(X_{n+1}|X_0, \dots, X_n)$ based on the observations (X_0, \dots, X_n) in a pointwise consistent way. It is well known that this is not possible at all values of n . We will estimate it along stopping times.

APPEARED IN: THEORY STOCH. PROCESS. 11 (2005), NO. 1-2, 112–120.

INTRODUCTION AND STATEMENT OF RESULTS

Suppose the distribution of the real-valued stationary time series $\{X_n\}_{n=0}^\infty$ is not known a priori. The goal is to estimate the conditional expectation $E(X_{n+1}|X_0, \dots, X_n)$ from the data segment X_0, \dots, X_n such that the difference between the estimate and the conditional expectation should tend to zero almost surely as the number of observations n tends to infinity. This problem (for binary time series) was introduced in Cover (1975). When one is obliged to estimate for all n , Bailey (1976) and Ryabko (1988) proved the nonexistence of such a universal algorithm even over the class of all stationary and ergodic binary time series.

In a special case, for certain Gaussian processes, Schäfer (2002) constructed an algorithm which can estimate the conditional expectation for every time instance n .

For further reading on related topics cf. Ornstein (1978), Algoet (1992), (1999), Morvai Yakowitz and Algoet (1997), Morvai, Yakowitz and Györfi (1996), Györfi, Lugosi and Morvai (1999), Györfi and Lugosi (2002), Weiss (2000) and Györfi et al. (2002).

In this paper we do not require to estimate for every time instance n , but rather, merely along a sequence of stopping times. That is, looking at the data segment X_0, \dots, X_n our rule will decide if we estimate for this n or not, but anyhow we will definitely estimate for infinitely many n . Algorithms of this kind were proposed for binary time series in Morvai (2003) and Morvai and Weiss (2003).

We will consider two-sided real-valued processes $\{X_n\}_{n=-\infty}^\infty$. A one-sided stationary time series $\{X_n\}_{n=0}^\infty$ can always be considered to be a two-sided stationary time series $\{X_n\}_{n=-\infty}^\infty$.

Let \mathfrak{R} be the set of all real numbers and put \mathfrak{R}^{*-} the set of all one-sided sequences of real numbers, that is,

$$\mathfrak{R}^{*-} = \{(\dots, x_{-1}, x_0) : x_i \in \mathfrak{R} \text{ for all } -\infty < i \leq 0\}.$$

1991 *Mathematics Subject Classification.* 62G05, 60G25, 60G10.

Key words and phrases. Nonparametric estimation, stationary processes.

Define the metric $d^*(\cdot, \cdot)$ on \mathfrak{R}^{*-} as

$$d^*((\dots, x_{-1}, x_0), (\dots, y_{-1}, y_0)) = \sum_{i=0}^{\infty} 2^{-i-1} \frac{|x_{-i} - y_{-i}|}{1 + |x_{-i} - y_{-i}|}.$$

Definition.: The conditional expectation $E(X_1 | \dots, X_{-1}, X_0)$ is almost surely continuous if for some set $B \subseteq \mathfrak{R}^{*-}$ which has probability one the conditional expectation $E(X_1 | \dots, X_{-1}, X_0)$ restricted to this set B is continuous with respect to metric $d^*(\cdot, \cdot)$.

Now we introduce our algorithm. For notational convenience, let $X_m^n = (X_m, \dots, X_n)$, where $m \leq n$. Define the nested sequence of partitions $\{\mathcal{P}_k\}_{k=0}^{\infty}$ of the real line as follows. Let

$$\mathcal{P}_k = \{[i2^{-k}, (i+1)2^{-k}) : \text{for } i = 0, 1, -1, 2, -2, \dots\}.$$

Let $x \rightarrow [x]^k$ denote a quantizer that assigns to any point $x \in \mathfrak{R}$ the unique interval in \mathcal{P}_k that contains x . Let $[X_m^n]^k = ([X_m]^k, \dots, [X_n]^k)$.

We define the stopping times $\{\lambda_n\}$ along which we will estimate. Set $\lambda_0 = 0$. For $n = 1, 2, \dots$, define λ_n recursively. Let

$$\lambda_n = \lambda_{n-1} + \min\{t > 0 : [X_t^{\lambda_{n-1}+t}]^n = [X_0^{\lambda_{n-1}}]^n\}. \quad (1)$$

Note that $\lambda_n \geq n$ and it is a stopping time on $[X_0^\infty]^n$. Let $f_k : \mathcal{P}_k \rightarrow \mathfrak{R}$ denote a function that assigns to any cell $A \in \mathcal{P}_k$ a point in A . The n th estimate m_n is defined as

$$m_n = \frac{1}{n} \sum_{j=0}^{n-1} f_j([X_{\lambda_j+1}]^j). \quad (2)$$

Observe that m_n depends solely on $[X_0^{\lambda_n}]^n$. This estimator can be viewed as a sampled version of the predictor in Morvai, Yakowitz and Györfi (1996), Weiss (2000), Algoet (1999) and Györfi et al. (2002).

Define the time series $\{\tilde{X}_n\}_{n=-\infty}^0$ as

$$\tilde{X}_{-n} = \lim_{j \rightarrow \infty} X_{\lambda_j - n} \text{ for } n \geq 0, \quad (3)$$

where the limit exists since the intervals $\{[X_{\lambda_j - n}]^j\}_{j=n}^{\infty}$ are nested and their lengths tend to zero.

Define the function $e : \mathfrak{R}^{*-} \rightarrow (-\infty, \infty)$ as

$$e(x_{-\infty}^0) = E(X_1 | X_{-\infty}^0 = x_{-\infty}^0).$$

We will prove the following theorem.

Theorem. *Let $\{X_n\}$ be a real-valued stationary time series with $E(|X_0|^2) < \infty$. Then almost surely*

$$\lim_{n \rightarrow \infty} m_n = \lim_{n \rightarrow \infty} E(X_{\lambda_n+1} | [X_0^{\lambda_n}]^n) = e(\tilde{X}_{-\infty}^0)$$

and

$$\lim_{n \rightarrow \infty} \left| m_n - E(X_{\lambda_n+1} | [X_0^{\lambda_n}]^n) \right| = 0.$$

Moreover, if in addition the conditional expectation $E(X_1 | X_{-\infty}^0)$ is almost surely continuous then almost surely

$$\lim_{n \rightarrow \infty} \left| m_n - E(X_{\lambda_n+1} | X_0^{\lambda_n}) \right| = 0.$$

Unfortunately, there is a stationary and ergodic Markov chain $\{X_n\}$ taking values from a countable subset of the unit interval such that

$$P \left(\limsup_{n \rightarrow \infty} \left| m_n - E(X_{\lambda_n+1} | X_0^{\lambda_n}) \right| > 0 \right) > 0.$$

Remarks.

Let $\{X_n\}$ be a real-valued stationary time series with $E(|X_0|^2) < \infty$. If the distribution of X_0 happens to concentrate on finitely many atoms then

$$E(X_{\lambda_n+1} | [X_0^{\lambda_n}]^n) = E(X_{\lambda_n+1} | X_0^{\lambda_n}) \text{ eventually}$$

and so $|m_n - E(X_{\lambda_n+1} | X_0^{\lambda_n})| \rightarrow 0$ almost surely, without any continuity condition.

Let $\{X_n\}$ be a real-valued stationary time series with $E(|X_0|^2) < \infty$. If one knows in advance that the distribution of X_0 concentrates on finite or countably infinite atoms then one may omit the partition \mathcal{P}_k , the quantizer $[\cdot]^k$ and the function $f_k(\cdot)$ entirely. That is, one may define $\lambda'_0 = 0$ and for $n = 1, 2, \dots$ set

$$\lambda'_n = \lambda'_{n-1} + \min\{t > 0 : X_t^{\lambda'_{n-1}+t} = X_0^{\lambda'_{n-1}}\}$$

and

$$m'_n = \frac{1}{n} \sum_{j=0}^{n-1} X_{\lambda'_j+1}.$$

Then

$$\lim_{n \rightarrow \infty} \left| m'_n - E(X_{\lambda'_n+1} | X_0^{\lambda'_n}) \right| = 0 \text{ almost surely}$$

without any continuity condition. Particularly, m'_n works for the counterexample process in the third part of the Theorem.

The counterexample Markov chain in the third part of the Theorem of course will not possess almost surely continuous conditional expectation $E(X_1 | X_{-\infty}^0)$.

From the proof of Bailey (1976), Ryabko (1988), Györfi, Morvai, Yakowitz (1998) it is clear that even for the class of all stationary and ergodic binary time series with almost surely continuous conditional expectation $E(X_1 | X_{-\infty}^0)$ one can not estimate $E(X_{n+1} | X_0^n)$ for all n in a pointwise consistent way.

PROOFS

It will be useful to define other processes $\{\hat{X}_n^{(k)}\}_{n=-\infty}^{\infty}$ for $k \geq 0$ as follows. Let

$$\hat{X}_{-n}^{(k)} = X_{\lambda_k - n} \text{ for } -\infty < n < \infty. \quad (4)$$

For an arbitrary real-valued stationary time series $\{Y_n\}$, let $\hat{\lambda}_0(Y_{-\infty}^0) = 0$ and for $n \geq 1$ define

$$\hat{\lambda}_n(Y_{-\infty}^0) = \hat{\lambda}_{n-1}(Y_{-\infty}^0) - \min\{t > 0 : [Y_{\hat{\lambda}_{n-1}-t}^{-t}]^n = [Y_{\hat{\lambda}_{n-1}}^0]^n\}.$$

Let T denote the left shift operator, that is, $(Tx_{-\infty}^\infty)_i = x_{i+1}$. It is easy to see that if $\lambda_n(x_{-\infty}^\infty) = l$ then $\hat{\lambda}_n(T^l x_{-\infty}^\infty) = -l$.

Proof of the Theorem.

Step 1. We show that for arbitrary $k \geq 0$, the time series $\{\hat{X}_n^{(k)}\}_{n=-\infty}^{\infty}$ and $\{X_n\}_{n=-\infty}^{\infty}$ have identical distribution.

It is enough to show that for all $k \geq 0$, $m \geq n \geq 0$, and Borel set $F \subseteq \mathbb{R}^{n+1}$,

$$P((\hat{X}_{m-n}^{(k)}, \dots, \hat{X}_m^{(k)}) \in F) = P(X_{m-n}^m \in F).$$

This is immediate by stationarity of $\{X_n\}$ and by the fact that for all $k \geq 0$, $m \geq n \geq 0$, $l \geq 0$, $F \subseteq \mathbb{R}^{n+1}$,

$$T^l\{X_{\lambda_k+m-n}^{\lambda_k+m} \in F, \lambda_k = l\} = \{X_{m-n}^m \in F, \hat{\lambda}_k(X_{-\infty}^0) = -l\}.$$

Step 2. We show that for $k \geq 0$, almost surely,

$$\hat{\lambda}_k(\dots, \hat{X}_{-1}^{(k)}, \hat{X}_0^{(k)}) = \hat{\lambda}_k(\tilde{X}_{-\infty}^0)$$

and

$$[\tilde{X}_{\hat{\lambda}_k(\tilde{X}_{-\infty}^0)}^0]^{k+1} = [\hat{X}_{\hat{\lambda}_k(\dots, \hat{X}_{-1}^{(k)}, \hat{X}_0^{(k)})}^{(k)}, \dots, \hat{X}_0^{(k)}]^{k+1}.$$

Since we are dealing with a nested sequence of partitions and $\hat{\lambda}_k$ depends solely on the k th quantized sequence, it is enough to prove that for any $i \geq 0$ and for all $j \geq i$, almost surely, $[\tilde{X}_{-i}]^{j+1} = [\hat{X}_{-i}^{(j)}]^{j+1}$. (Note that $\lambda_j(X_0^\infty) - j \geq 0$.) If $\tilde{X}_{-i} \notin [\hat{X}_{-i}^{(j)}]^{j+1}$ for some $j \geq i$ then this must happen at a right end-point of some interval in $\bigcup_{k=0}^{\infty} \mathcal{P}_k$. By (3) and Step 1, we have

$$\begin{aligned} & 1 - P(\tilde{X}_{-i} \in [\hat{X}_{-i}^{(j)}]^{j+1} \text{ for all } j \geq i) \\ & \leq \sum_{k=i}^{\infty} \sum_{s=-\infty}^{\infty} P(\tilde{X}_{-i} = s2^{-k}, \hat{X}_{-i}^{(j)} < \tilde{X}_{-i} \text{ for all } j \geq k) \\ & \leq \sum_{k=i}^{\infty} \sum_{s=-\infty}^{\infty} \lim_{j \rightarrow \infty} P(s2^{-k} - 2^{-j} \leq \hat{X}_{-i}^{(j)} < s2^{-k}) \\ & = \sum_{k=i}^{\infty} \sum_{s=-\infty}^{\infty} \lim_{j \rightarrow \infty} P(s2^{-k} - 2^{-j} \leq X_{-i} < s2^{-k}) \\ & = 0. \end{aligned}$$

Step 3. We show that the distributions of $\{\tilde{X}_n\}_{n=-\infty}^0$ and $\{X_n\}_{n=-\infty}^0$ are the same.

This is immediate from Step 1 and Step 2.

The time series $\{\tilde{X}_n\}_{n=-\infty}^0$ is stationary, since $\{X_n\}_{n=-\infty}^0$ is stationary, and it can be extended to be a two-sided time series $\{\tilde{X}_n\}_{n=-\infty}^\infty$. We will use this fact only for the purpose of defining the conditional expectation $E(\tilde{X}_1|\tilde{X}_{-\infty}^0)$.

Step 4. We prove the first part of the Theorem.

Consider

$$\begin{aligned} m_n &= \frac{1}{n} \sum_{j=0}^{n-1} \left(f_j([X_{\lambda_j+1}]^j) - E(f_j([X_{\lambda_j+1}]^j)|[X_0^{\lambda_j}]^j) \right) \\ &\quad + \frac{1}{n} \sum_{j=0}^{n-1} \left(E(f_j([X_{\lambda_j+1}]^j)|[X_0^{\lambda_j}]^j) - E(X_{\lambda_j+1}|[X_0^{\lambda_j}]^j) \right) \\ &\quad + \frac{1}{n} \sum_{j=0}^{n-1} E(X_{\lambda_j+1}|[X_0^{\lambda_j}]^j). \end{aligned} \quad (5)$$

Observe that $\{\Gamma_j = f_j([X_{\lambda_j+1}]^j) - E(f_j([X_{\lambda_j+1}]^j)|[X_0^{\lambda_j}]^j)\}$ is a sequence of orthogonal random variables with $E\Gamma_j = 0$ and $E(\Gamma_j^2) \leq E(|X_1|^2) + 2E|X_1| + 1$ since $E(\Gamma_j^2) \leq E(|X_{\lambda_j+1}|^2) + 2E|X_{\lambda_j+1}| + 1$ and, by Step 1, X_{λ_j+1} has the same distribution as X_1 . Now by Theorem 3.2.2 in Révész (1968),

$$\frac{1}{n} \sum_{j=0}^{n-1} \Gamma_j \rightarrow 0 \text{ almost surely.}$$

The second term tends to zero since $|f_j([X_{\lambda_j+1}]^j) - X_{\lambda_j+1}| \leq 2^{-j}$. Now we deal with the third term. By Step 2, Step 1 and Step 3,

$$E(X_{\lambda_j+1}|[X_0^{\lambda_j}]^j) = E(\tilde{X}_1|\tilde{X}_{\lambda_j(\tilde{X}_{-\infty}^0)}^j).$$

The latter forms a martingale and by Theorem 7.6.2 in Ash (1972), almost surely,

$$E(X_{\lambda_j+1}|[X_0^{\lambda_j}]^j) = E(\tilde{X}_1|\tilde{X}_{\lambda_j(\tilde{X}_{-\infty}^0)}^j) \rightarrow E(\tilde{X}_1|\tilde{X}_{-\infty}^0). \quad (6)$$

By (5) and (6), almost surely,

$$\lim_{n \rightarrow \infty} m_n = E(\tilde{X}_1|\tilde{X}_{-\infty}^0). \quad (7)$$

Thus the first part of the Theorem is proved.

Step 5. We prove the second part of the Theorem.

By (7) it is enough to prove that almost surely $E(X_{\lambda_j+1}|X_0^{\lambda_j}) \rightarrow E(\tilde{X}_1|\tilde{X}_{-\infty}^0)$ provided that $E(X_1|X_{-\infty}^0)$ is almost surely continuous. By assumption, the function $e(\cdot)$ is continuous on a set $B \subseteq \mathfrak{R}^{*-}$ with $P(X_{-\infty}^0 \in B) = 1$. By Step 1 and Step 3,

$$P(\tilde{X}_{-\infty}^0 \in B, (\dots, \hat{X}_{-1}^{(j)}, \hat{X}_0^{(j)}) \in B \text{ for all } j \geq 0) = 1. \quad (8)$$

Let

$$\mathcal{N}_j(X_0^{\lambda_j}) = \{z_{-\infty}^0 \in \mathfrak{R}^{*-} : z_{-\lambda_j} \in [X_0]^{(j)}, \dots, z_0 \in [X_{\lambda_j}]^{(j)}\}.$$

By (4), (8) and Step 2, almost surely, for all j ,

$$(\dots, \hat{X}_{-1}^{(j)}, \hat{X}_0^{(j)}) \in \mathcal{N}_j(X_0^{\lambda_j}) \cap B \text{ and } \tilde{X}_{-\infty}^0 \in \mathcal{N}_j(X_0^{\lambda_j}) \cap B. \quad (9)$$

Put

$$\Theta_j(X_0^{\lambda_j}) = \sup_{y_{-\infty}^0, z_{-\infty}^0 \in \mathcal{N}_j(X_0^{\lambda_j}) \cap B} |e(y_{-\infty}^0) - e(z_{-\infty}^0)|.$$

Since $e(\cdot)$ is continuous on set B and by (9), almost surely,

$$\lim_{j \rightarrow \infty} \Theta_j(X_0^{\lambda_j}) = 0. \quad (10)$$

By (9) and (10), almost surely,

$$\begin{aligned} & \limsup_{j \rightarrow \infty} \left| E\left(e(\tilde{X}_{-\infty}^0) | [X_0^{\lambda_j}]^{(j)}\right) - E\left(e(\dots, \hat{X}_{-1}^{(j)}, \hat{X}_0^{(j)}) | X_0^{\lambda_j}\right) \right| \\ & \leq \limsup_{j \rightarrow \infty} E\left(\left| E\left(e(\tilde{X}_{-\infty}^0) | [X_0^{\lambda_j}]^{(j)}\right) - e(\dots, \hat{X}_{-1}^{(j)}, \hat{X}_0^{(j)}) \right| | X_0^{\lambda_j}\right) \\ & \leq \limsup_{j \rightarrow \infty} E\left(\Theta_j(X_0^{\lambda_j}) | X_0^{\lambda_j}\right) \\ & = \limsup_{j \rightarrow \infty} \Theta_j(X_0^{\lambda_j}) \\ & = 0. \end{aligned} \quad (11)$$

By Step 2,

$$\begin{aligned} E\left(X_{\lambda_j+1} | X_0^{\lambda_j}\right) &= E\left(e(\tilde{X}_{-\infty}^0) | [\tilde{X}_{\lambda_j}^0]^{(j)}\right) \\ &\quad - \left\{ E\left(e(\tilde{X}_{-\infty}^0) | [X_0^{\lambda_j}]^{(j)}\right) - E\left(e(\dots, \hat{X}_{-1}^{(j)}, \hat{X}_0^{(j)}) | X_0^{\lambda_j}\right) \right\}. \end{aligned}$$

The first term tends to $e(\tilde{X}_{-\infty}^0)$ by the almost sure martingale convergence theorem (cf. Theorem 7.6.2 in Ash (1972)) since by Step 3, $E|e(\tilde{X}_{-\infty}^0)| \leq E|\tilde{X}_1| = E|X_1| < \infty$. The second term tends to zero by (11). The proof of the second part of the Theorem is complete.

Step 6. *We prove the third part of the Theorem.*

First we define a Markov chain $\{M_n\}$ on the nonnegative integers which will serve as a technical tool for our counterexample process. Let the transition probabilities be as follows.

$$P(M_1 = 0 | M_0 = 0) = P(M_1 = 1 | M_0 = 0) = P(M_1 = 0 | M_0 = 1) = 2^{-1}$$

and for $i = 2, 3, \dots$, let

$$P(M_1 = i | M_0 = 1) = 2^{-i} \text{ and } P(M_1 = 0 | M_0 = i) = 1.$$

All other transitions happen with probability zero. Note that one can reach state 1 only from state 0. It is easy to see that the Markov chain just defined yields a stationary and ergodic time series with initial probabilities $P(M_0 = 0) = \frac{4}{7}$, $P(M_0 = 1) = \frac{2}{7}$, and for $i = 2, 3, \dots$ $P(M_0 = i) = \frac{1}{7 \cdot 2^{i-1}}$. Our counterexample process $\{X_n\}$ will be a one to one function of the Markov chain $\{M_n\}$. Define the function $h : \{0, 1, 2, \dots\} \rightarrow \mathbb{R}$ as $h(0) = 0$, $h(1) = 1$ and for $i \geq 2$ put $h(i) = \frac{2^{-2^i}}{2}$. Let $X_n = h(M_n)$. Since $h(\cdot)$ is one to one, $\{X_n\}$ is also a Markov chain. Since $\{\tilde{X}_n\}$ has the same distribution as $\{X_n\}$, $\{\tilde{X}_n\}$ is also a Markov chain. Let

$$A_n = \{h(i) : h(i) < 2^{-(n+1)} \text{ for } i = 0, 1, 2, \dots\}.$$

Note that $h(i) \in A_n$ if and only if $[h(i)]^{n+1} = [0]^{n+1}$. Define the event

$$H = \{\tilde{X}_0 = 0, X_0^1 = (0, 1)\}.$$

Observe: If $X_1 = 1$ then $X_0 = 0$. (State 1 can be reached only from state 0.) The event $\{\tilde{X}_0 = 0\}$ happens if and only if $X_{\lambda_n} \in A_n$ for all $n = 1, 2, \dots$. Since $[h(0)]^1 = [h(i)]^1$ for $i \geq 2$ and for all $k \geq 0$, $[h(1)]^k \neq [h(i)]^k$ provided $i \neq 1$ the event $\{\tilde{X}_{-1} = 1\}$ occurs if and only if $X_1 = 1$. It follows that

$$H = \{X_0 = 0, X_1 = 1, X_{\lambda_n} \in A_n \text{ for } n = 1, 2, \dots\} = \{\tilde{X}_{-2}^0 = (0, 1, 0)\}.$$

Since the time series $\{\tilde{X}_n\}$ has the same distribution as $\{X_n\}$,

$$P(H) = P(X_{-2}^0 = (0, 1, 0)) = \frac{4}{7} \frac{1}{2} \frac{1}{2} = \frac{1}{7} > 0.$$

It will be enough to show that $X_{\lambda_n} \in A_n - \{0\}$ happens infinitely often given the condition H since if $X_{\lambda_n} \in A_n - \{0\}$ happens then $X_{\lambda_{n+1}} = 0$ and by (7), on H

$$m_n \rightarrow E(\tilde{X}_1 | \tilde{X}_0 = 0) = 0.5$$

and so

$$P\left(\limsup_{n \rightarrow \infty} |m_n - E(X_{\lambda_{n+1}} | X_0^{\lambda_n})| = 0.5 | H\right) = 1$$

and $P(H) > 0$. To prove that $\{X_{\lambda_n} \in A_n - \{0\}\}$ occurs infinitely often we need the following observation for repeated use: By the Markov property and the construction in (1) if $x_i \in A_i$ for $i = 1, 2, \dots, j$ then for $j \geq 1$,

$$P(X_{\lambda_j} = x_j | X_0^1 = (0, 1), X_{\lambda_m} = x_m \text{ for } 1 \leq m < j) = P(X_1 = x_j | X_0 = 1, X_1 \in A_{j-1}). \quad (12)$$

Indeed, for $j = 1$ this is trivial, since $X_1 = 1$ implies that $X_0 = 0$, $\lambda_1 = 2$ while $X_0 = 1$ implies that $X_1 \in A_0$. For $j \geq 2$ set $\psi_0^j = \lambda_{j-1} - 1$ and for $i \geq 1$ the ψ_i^j will be the successive occurrences of the block $[X_0^{\lambda_{j-1}-1}]^j$ in the j -th quantization, defined by

$$\psi_i^j = \min\{t > \psi_{i-1}^j : [X_{t-\lambda_{j-1}+1}^t]^j = [X_{\psi_{i-1}^j-\lambda_{j-1}+1}^{\psi_i^j}]^j\}.$$

These ψ_i^j are stopping times for $i = 1, 2, \dots$. Temporarily let D_j denote the event

$$\{X_0^1 = (0, 1), X_{\lambda_m} = x_m \text{ for } 1 \leq m < j\}.$$

The way that λ_j is defined means that on D_j if λ_j occurs at the i -th repetition of $[X_0^{\lambda_{j-1}-1}]^j$ it is because $\psi_i^j < \lambda_j$ and $X_{\psi_i^j+1} \in A_{j-1}$. It follows that

$$P(X_{\lambda_j} = x_j | D_j) = \sum_{i=1}^{\infty} P(X_{\psi_i^j+1} = x_j | X_{\psi_i^j+1} \in A_{j-1}, \psi_i^j < \lambda_j, D_j) P(\psi_i^j + 1 = \lambda_j | D_j).$$

Since $x_j \in A_j \subseteq A_{j-1}$, each expression $P(X_{\psi_i^j+1} = x_j | X_{\psi_i^j+1} \in A_{j-1}, \psi_i^j < \lambda_j, D_j)$ can be written as

$$P(X_{\psi_i^j+1} = x_j | X_{\psi_i^j+1} \in A_{j-1}, \psi_i^j < \lambda_j, D_j) = \frac{P(X_{\psi_i^j+1} = x_j | \psi_i^j < \lambda_j, D_j)}{P(X_{\psi_i^j+1} \in A_{j-1} | \psi_i^j < \lambda_j, D_j)}$$

and then by decomposition according to the value l of ψ_i^j we get

$$\begin{aligned} & P(X_{\psi_i^j+1} = x_j | \psi_i^j < \lambda_j, D_j) \\ &= \sum_{l=1}^{\infty} \left(\frac{P(X_{l+1} = x_j | \psi_i^j = l < \lambda_j, D_j)}{P(X_{l+1} \in A_{j-1} | \psi_i^j = l < \lambda_j, D_j)} P(\psi_i^j = l, X_{\psi_i^j+1} \in A_{j-1} | \psi_i^j < \lambda_j, D_j) \right). \end{aligned}$$

Observe that $X_{\psi_i^j} = 1$ provided $X_1 = 1$ and the event $\{\psi_i^j < \lambda_j\}$ is measurable with respect to $\sigma([X_0^{\psi_i^j}]^j)$. Now by the Markov property we get

$$\begin{aligned} & P(X_{\psi_i^j+1} = x_j | X_{\psi_i^j+1} \in A_{j-1}, \psi_i^j < \lambda_j, D_j) \\ &= \sum_{l=1}^{\infty} \left(\frac{P(X_{l+1} = x_j | X_l = 1)}{P(X_{l+1} \in A_{j-1} | X_l = 1)} \cdot \frac{P(\psi_i^j = l, X_{\psi_i^j+1} \in A_{j-1} | \psi_i^j < \lambda_j, D_j)}{P(X_{\psi_i^j+1} \in A_{j-1} | \psi_i^j < \lambda_j, D_j)} \right). \end{aligned}$$

By stationarity and since $x_j \in A_j \subseteq A_{j-1}$,

$$\frac{P(X_{l+1} = x_j | X_l = 1)}{P(X_{l+1} \in A_{j-1} | X_l = 1)} = P(X_1 = x_j | X_1 \in A_{j-1}, X_0 = 1).$$

Combining all this we get

$$\begin{aligned} & P(X_{\lambda_j} = x_j | D_j) \\ &= P(X_1 = x_j | X_1 \in A_{j-1}, X_0 = 1) \\ &\cdot \left(\sum_{i=1}^{\infty} P(\psi_i^j + 1 = \lambda_j | D_j) \sum_{l=1}^{\infty} \frac{P(\psi_i^j = l, X_{\psi_i^j+1} \in A_{j-1} | \psi_i^j < \lambda_j, D_j)}{P(X_{\psi_i^j+1} \in A_{j-1} | \psi_i^j < \lambda_j, D_j)} \right) \\ &= P(X_1 = x_j | X_1 \in A_{j-1}, X_0 = 1) \sum_{i=1}^{\infty} P(\psi_i^j + 1 = \lambda_j | D_j) \\ &= P(X_1 = x_j | X_1 \in A_{j-1}, X_0 = 1) \end{aligned}$$

and we have proved (12).

In order to show that the events

$$\{X_{\lambda_n} \in A_n - \{0\}\}$$

occur infinitely often we prove that they have sufficiently large conditional probabilities and they are conditionally independent given the condition H . First we calculate $P(X_{\lambda_n} \in A_n - \{0\}|H)$. For $n \geq 2$, by (12),

$$\begin{aligned} & P(X_{\lambda_n} \in A_n - \{0\}|H) \\ &= \frac{P(\{X_{\lambda_n} \in A_n - \{0\}\} \cap H)}{P(H)} \\ &= \frac{P(X_{\lambda_n} \in A_n - \{0\} | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < n)}{P(X_{\lambda_n} \in A_n | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < n)} \\ &\cdot \prod_{m=n+1}^{\infty} \frac{P(X_{\lambda_m} \in A_m | X_0^1 = (0, 1), X_{\lambda_n} \in A_n - \{0\}, X_{\lambda_j} \in A_j \text{ for } 1 \leq j < m)}{P(X_{\lambda_m} \in A_m | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < m)} \\ &= \frac{P(X_{\lambda_n} \in A_n - \{0\} | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < n)}{P(X_{\lambda_n} \in A_n | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < n)} \\ &\geq P(X_{\lambda_n} \in A_n - \{0\} | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < n) \\ &= P(X_1 \in A_n, X_1 \neq 0 | X_0 = 1, X_1 \in A_{n-1}) \\ &\geq P(X_1 \in A_n, X_1 \neq 0 | X_0 = 1) \\ &= \sum_{i \in A_n - \{0\}} \frac{1}{2^i} \\ &= \sum_{i > \log_2(n)} \frac{1}{2^i} \\ &\geq \frac{1}{n}. \end{aligned}$$

We have just proved that

$$\sum_n P(X_{\lambda_n} \in A_n - \{0\}|H) \geq \sum_n \frac{1}{n} = \infty. \quad (13)$$

Now we will prove that for $n = 1, 2, \dots$, the events $\{X_{\lambda_n} \in A_n - \{0\}\}$ are conditionally independent given H . Since

$$\begin{aligned} & P(X_{\lambda_i} \in A_i - \{0\} \text{ for } i = 1, 2, \dots, k | H) \\ &= \sum_{x_1 \in A_1 - \{0\}} \cdots \sum_{x_k \in A_k - \{0\}} P(X_{\lambda_i} = x_i \text{ for } i = 1, 2, \dots, k | H) \end{aligned}$$

it is enough to show that the events $\{X_{\lambda_i} = x_i\}$ are conditionally independent given the

condition H , provided that $x_i \in A_i$. Let $x_i \in A_i$. Then by repeated use of (12)

$$\begin{aligned}
& P(X_{\lambda_i} = x_i \text{ for } i = 1, 2, \dots, k | H) \\
&= \frac{P(X_{\lambda_i} = x_i \text{ for } i = 1, 2, \dots, k, H)}{P(H)} \\
&= \left(\prod_{m=1}^k \frac{P(X_{\lambda_m} = x_m | X_0^1 = (0, 1), X_{\lambda_j} = x_j \text{ for } 1 \leq j < m)}{P(X_{\lambda_m} \in A_m | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < m)} \right) \\
&\cdot \prod_{l=k+1}^{\infty} \frac{P(X_{\lambda_l} \in A_l | X_0^1 = (0, 1), X_{\lambda_i} = x_i \text{ for } 1 \leq i \leq k \text{ and } X_{\lambda_j} \in A_j \text{ for } 1 \leq j < l)}{P(X_{\lambda_l} \in A_l | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < l)} \\
&= \prod_{m=1}^k \frac{P(X_{\lambda_m} = x_m | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < m)}{P(X_{\lambda_m} \in A_m | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < m)} \\
&= \prod_{m=1}^k \left(\frac{P(X_{\lambda_m} = x_m | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < m)}{P(X_{\lambda_m} \in A_m | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < m)} \right. \\
&\cdot \left. \prod_{l=m+1}^{\infty} \frac{P(X_{\lambda_l} \in A_l | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < l)}{P(X_{\lambda_l} \in A_l | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < l)} \right) \\
&= \prod_{m=1}^k \left(\frac{P(X_{\lambda_m} = x_m | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < m)}{P(X_{\lambda_m} \in A_m | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < m)} \right. \\
&\cdot \left. \prod_{l=m+1}^{\infty} \frac{P(X_{\lambda_l} \in A_l | X_0^1 = (0, 1), X_{\lambda_m} = x_m, X_{\lambda_j} \in A_j \text{ for } 1 \leq j < l)}{P(X_{\lambda_l} \in A_l | X_0^1 = (0, 1), X_{\lambda_j} \in A_j \text{ for } 1 \leq j < l)} \right) \\
&= \prod_{i=1}^k \frac{P(X_{\lambda_i} = x_i, H)}{P(H)} \\
&= \prod_{i=1}^k P(X_{\lambda_i} = x_i | H).
\end{aligned}$$

Now by (13) and the Borel-Cantelli lemma (cf. Lemma B in Rényi (1970) on page 390) the events $\{X_{\lambda_n} \in A_n - \{0\}\}$ occur infinitely often and the third part of the Theorem is proved. The proof of the Theorem is complete.

REFERENCES

1. P. Algoet, *Universal schemes for prediction, gambling and portfolio selection*, Annals of Probability **20** (1992), 901–941.
2. P. Algoet, *Universal schemes for learning the best nonlinear predictor given the infinite past and side information*, IEEE Transactions on Information Theory **45** (1999), no. 4, 1165–1185.
3. R.B. Ash, *Real Analysis and Probability*, “Academic Press”, New York, 1972.
4. D. H. Bailey, *Sequential Schemes for Classifying and Predicting Ergodic Processes*, Ph. D. thesis, “Stanford University”, 1976.
5. T. M. Cover, *Open problems in information theory*, In: 1975 IEEE Joint Workshop on Information Theory, “IEEE Press”, New York, 1975, pp. 35–36.
6. L. Györfi, M. Kohler, A. Krzyżak, and H. Walk, *A Distribution Free Theory of Nonparametric Regression*, “Springer-Verlag”, New York, 2002.
7. L. Györfi and G. Lugosi, *Strategies for sequential prediction of stationary time series*, Modeling Uncertainty An Examination of Stochastic Theory, Methods, and Applications M.Dror, P. L’Ecuyer, F. Szidarovszky (Eds.), “Kluwer Academic Publishers”, 2002, pp. 225–248.

8. L. Györfi, G. Lugosi and G. Morvai, *A simple randomized algorithm for consistent sequential prediction of ergodic time series*, IEEE Transactions on Information Theory **45** (1999), no. 45, 2642–2650.
9. L. Györfi, G. Morvai, and S. Yakowitz, *Limits to consistent on-line forecasting for ergodic time series*, IEEE Transactions on Information Theory **44** (1998), no. 2, 886–892.
10. G. Morvai, *Guessing the output of a stationary binary time series*, in: Foundations of Statistical Inference Y. Haitovsky, H.R. Lerche, Y. Ritov (Eds.) (2003), “Physika Verlag”, 205–213.
11. G. Morvai and B. Weiss, *Forecasting for stationary binary time series*, Acta Applicandae Mathematicae **79** (2003), no. 1-2,, 25–34.
12. G. Morvai, S. Yakowitz, and P. Algoet, *Weakly convergent nonparametric forecasting of stationary time series*, IEEE Transactions on Information Theory **43** (1997), no. 2, 483–498.
13. G. Morvai, S. Yakowitz, and L. Györfi, *Nonparametric inferences for ergodic, stationary time series*, Annals of Statistics **24** (1996), no. 1, 370–379.
14. D. S. Ornstein, *Guessing the next output of a stationary process*, Israel J. Math **30** (1978), 292–296.
15. A. Rényi, *Probability Theory*, “Akadémiai Kiadó”, 1970.
16. P. Révész, *The Law of Large Numbers*, “Academic Press”, 1968.
17. B. Ya. Ryabko, *Prediction of random sequences and universal coding*, Problems of Inform. Trans. (Problemy Peredachi Informatsii) **24** (1988), no. 2, 3–14.
18. D. Schäfer, *Strongly consistent online forecasting of centered Gaussian processes*, IEEE Transactions on Information Theory **48** (2002), no. 3, 791–799.
19. B. Weiss, *Single Orbit Dynamics*, “American Mathematical Society”, Providence, RI, 2000.

GUSZTÁV MORVAI (CORRESPONDING AUTHOR. TEL.: 36-1-4632867; FAX.:36-1-4633147.) RESEARCH GROUP FOR INFORMATICS AND ELECTRONICS OF THE HUNGARIAN ACADEMY OF SCIENCES, BUDAPEST, 1521 GOLDMANN GYÖRGY TÉR 3, HUNGARY
E-mail address: `morvai@math.bme.hu`

BENJAMIN WEISS (TEL.: 972-2-658-4388; FAX.: 972-2-563-0702.) HEBREW UNIVERSITY OF JERUSALEM JERUSALEM 91904 ISRAEL
E-mail address: `weiss@math.huji.ac.il`